PB82-154972

Environmental Assessment of Industrial Discharges Based on Multiplicative Models

(U.S.) Industrial Environmental Research Lab.
Research Triangle Park, NC

16 Jun 81

U.S. Department of Commerce
National Technical Information Service
NTIS

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before comple

1. REPORT NO. EPA-600/J-81-490	2. Journal Article
3	PB82-154972
4. TITLE AND SUBTITLE: Environmental Assessment of Industrial Discharges Based on Multiplicative Models	5. REPORT DATE: December 1981
	6. PERFORMING ORGANIZATION CODE: 026079
7. AUTHOR(S): M. Ross Leadbetter and W. Gene Tucker (EPA/IERL-RTP)	8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS: see block 12	10. PROGRAM ELEMENT NO.
	11. CONTRACT/GRANT NO.
12. SPONSORING AGENCY NAME AND ADDRESS: **EPA, Office of Research and Development, Industrial Environmental Research Laboratory, Research Triangle Park, NC 27711**	13. TYPE OF REPORT AND PERIOD COVERED: Journal Article; 1981
	14. SPONSORING AGENCY CODE: **EPA/600/13**

15. SUPPLEMENTARY NOTES: Published in Environmental Science and Technology, Vol. 15, No. 11, pp 1355-1360, 1981.

16. ABSTRACT

The paper describes a method of environmental assessment of industrial discharges based on multiplicative models. The severity, S, of a substance in an industrial discharge is defined as the ratio of substance concentration, either at the source or at some ambient point of interest, to a maximum specified "safe" concentration level. The source is considered "clean" unless S is expected to exceed unity on more than a given acceptably small proportion of time. Otherwise, it is "dirty". A source is classified as clean or dirty based on (a) measurements of factors such as stack emission characteristics and (b) possible knowledge of the statistical properties of other factors (such as meteorology). Standard statistical decision techniques are used, with some novelty, to take account of the forms of variation present (time fluctuations, measurement errors, etc.) and to best incorporate existing prior knowledge of the statistical parameters involved. Log-normal distributional assumptions are used, coupled with multiplicative transport models in ambient cases.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
Pollution	Pollution Control	13B
Assessments	Stationary Sources	14B
Effluents	Environmental Assessment	
Flue Gases	Multiplicative Models	21B
Statistics		12A

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA 22161

18. DISTRIBUTION STATEMENT: **Release to Public**	19. SECURITY CLASS *(This Report)*: **Unclassified**	21. NO. OF PAGES: 6
	20. SECURITY CLASS *(This page)*: **Unclassified**	22. PRICE

EPA Form 2220-1 (9-73)

EPA-600/J-81-490
Journal Article

Environmental Assessment of Industrial Discharges Based on Multiplicative Models

M. Ross Leadbetter*† and W. Gene Tucker‡

U.S. Environmental Protection Agency, Industrial Environmental Research Laboratory, Research Triangle Park, North Carolina 27711

■ The *severity*, S, of a substance in a discharge from an industrial source is defined as the ratio of substance concentration, either at the source or at some ambient point of interest, to a maximum specified "safe" concentration level. The source is considered "clean" unless S is expected to exceed unity on more than a given acceptably small proportion of time. Otherwise, it is "dirty". The classification of a source as clean or dirty is made from (a) measurements of factors such as stack emission characteristics and (b) possible knowledge of the statistical properties of other factors (such as meteorology). Standard statistical decision techniques are used, with some novelty to take account of the forms of variation present (time fluctuations, measurement errors, etc.) and to best incorporate existing prior knowledge of the statistical parameters involved. Log normal distributional assumptions are used, coupled with multiplicative transport models in ambient cases.

I. Introduction

A current approach to assessment of substances discharged from industrial processes (*1*, *2*) is to compare the concentration, C, of the substance at some point of interest (either at the point of discharge or at some downstream point) with some estimated "safe" concentration level, or standard, G. The *severity* of the source is then defined as the ratio $S = C/G$.

Suppose, for the moment, that there are no temporal (e.g., daily) fluctuations in a source's discharge and ambient conditions. Then S has a constant value, and the source may be termed "clean" or "dirty" depending upon whether this value is less or greater than 1 (i.e., $C < G$ or $C > G$). The terms "clean" and "dirty" have no official standing, of course, but are conveniently brief and clear for our purposes here. S, of course, is unknown, but can be *estimated* from measurements of appropriate factors (e.g., emissions) giving some calculated estimate $\hat{S}$. The classification of the source as clean or dirty would be made on the basis of whether $\hat{S}$ was smaller or larger than some critical level, c (not to be confused with C denoting concentration). Since the measured factors (hence, also $\hat{S}$) are subject to measurement errors, it is intuitively plausible that, to have small probability of misclassifying dirty sources, c must be less than 1. Its precise value would be determined by standard statistical (hypothesis testing) techniques in such a way that the probability of misclassifying a dirty source is limited to some preassigned small value, β (e.g., 0.05).

If the severity at the discharge point is of primary interest, $\hat{S}$ will be calculated directly from measurements of concentration at that point. On the other hand, when *ambient* severities are of concern, an appropriate transport (e.g., diffusion) model must be assumed in order to relate C and S to

† Under Loan Agreement with University of North Carolina, Statistics Department, Chapel Hill, NC.
‡ Chief, Special Studies Staff, Industrial Environmental Research Laboratory.

This article not subject to U.S. Copyright. Published 1981 American Chemical Society

the factors which will be measured at the source site. Certain useful models—of *multiplicative* type—are especially convenient in dealing with the necessary underlying statistical calculations. These will be discussed in section II.

The foregoing discussion assumed that discharged amounts (hence, S) did not change with time. In reality, of course, S will have time variation (including a random component) and is likely—however small it may usually be—to at least occasionally exceed 1 (i.e., the concentration C will then exceed the safe level G). It is natural, therefore, to modify the definition of clean and dirty sources as follows. A source will be termed dirty if S is expected to exceed 1 on more than some specified (small) proportion, α, of time. Otherwise, of course, S will be termed clean. The number α may be of the order of 5% (or 1%) and will be chosen by considerations of the health hazard produced when S exceeds 1. This type of criterion is in current use for certain pollutants (e.g., ref *3*).

Again, the classification of a source as clean or dirty is a statistical decision problem, based on the value of an "estimated severity", $\hat{S}$, to be calculated from measurements made. Section III discusses the sources of variability, and section IV the classification procedures and statistical assumptions on which they are based. These primarily involve standard statistical techniques, with some novel aspects to properly take account of the different types of variability involved. Further, as discussed later, it is not necessary to measure factors (as perhaps wind speed) whose statistical properties are known from experience, and this knowledge may be utilized to improve the procedure. Section V is concerned with the same topic when multiple measurements are made.

The classification procedure limits the probability of misclassifying a dirty source, to some specified value, and in this way protects environmental interests. The probability of misclassification of clean sources (possibly causing unnecessary expense in discharge reduction) is also of interest. This topic will be taken up in section VI. The procedures described are illustrated in section VII, by an example involving actual effluent measurements on a textile weaving plant. The mathematical details of the construction of the procedure and for estimation of the variability parameters are given in Appendixes A and B.

II. Multiplicative Models

Typically, measurements are made at the point of discharge, and our interest lies in the severity either there or at some ambient point. In the latter case, as noted above, some form of transport model must be used to relate the ambient concentration, C, to appropriate discharge and other factors; i.e., a relationship

$$C = f(U,V,W,\ldots)$$

is used, where U, V, W, ... are the relevant source and ambient factors. For air pollution this model may involve the use of a "Gaussian plume", whereas for receiving bodies of water it may be based on relatively simple mixing considerations.

In many important cases the models used are either multiplicative or approximately so, in the sense that they involve solely products and quotients of the relevant variables. Two common examples follow.

Example 1. Air Pollution from an Industrial Stack. In this case a Gaussian plume dispersion model may be used under appropriate conditions, leading to the maximum ground-level concentration, C, of the substance of concern given by the formula

$$C = kUW/(VZ^2) \qquad (1)$$

where k is a constant, U = emission factor (pollutant mass per unit mass of product), W = production capacity (rate of production of product), V = wind velocity, and Z = effective stack plume height.

Example 2. Water Pollution from an Industrial Effluent. If an effluent is discharged at some point into a stream, simple geometrical considerations show that its concentration, sufficiently far downstream to ensure thorough mixing, is given by

$$C = UV/W \qquad (2)$$

where U = concentration of substance in the effluent, V = effluent discharge rate, and W = river flow rate. Note that various simplifying assumptions are made in such a model. In particular it is assumed that the effluent discharge rate is relatively small in comparison with the river flow rate. This assumption is reasonable for many situations and leads to the stated purely multiplicative form of the model.

As previously noted, some of the factors in such models (e.g., discharge factors) will be measured whereas others (such as wind speed), whose statistical properties are known from past records, may not. We may therefore write a general multiplicative model as

$$C = XY \qquad (S = XY/G) \qquad (3)$$

where X is the product–quotient of the measured factors, and Y that of the remaining factors which are not measured. For instance, example 1 above may have $X = UW$, $Y = k(VZ^2)$, and example 2, $X = UV$, $Y = 1/W$.

Finally, it will be convenient to use (here and throughout) the notation A^* to denote the (natural) logarithm of a number A, ($A^* = \ln A$). Using this notation, the general model (*3*) may be written additively:

$$C^* = X^* + Y^*$$

or

$$S^* = X^* + Y^* - G^*$$

III. Sources of Variability

The following sources of variation are relevant to this discussion: (1) time (e.g., daily) variation of the factors U, V, W, ... and hence of X, Y, and $S = XY/G$; (2) measurement errors in measured factors, i.e., those of which X is comprised; (3) uncertainty in the value of the safe level, G; (4) uncertainty in the "fit" of the model. The main concern here is with the temporal and measurement-error variation, 1 and 2, so that it is convenient to assume that G is known precisely and the error in the model is not significant. If G is not known, but obtained from, e.g., epidemiological experiments, it may well have readily obtained statistical properties which may be included simply in the analysis. Lack of fit of the model may also be simply accounted for, if desired, by inclusion of a reasonable random error term.

The detailed assumptions regarding variations 1 and 2 will be dealt with in Appendix A. However, note, as regards variation 1, that the basic statistical assumptions to be made are that the relevant factors U, V, W, ... have independent log normal distributions on a given day and that their values on different days are independent. The log normal nature of the factors is well documented in many cases. The independence is clear in some cases and requires individual checking in others. It is clear in principle that some statistical dependence (correlation) between values on different days may be included in the analysis, but this has not been investigated in detail.

In addition to the above assumptions concerning the temporal variability of the factors, any *measurements* of a factor will contain an additional measurements error (as noted under

the source of variation 2, above). Specifically, $\hat{U}$ denotes the *measured* value of a factor U, and $\hat{U} = \epsilon U$, where ϵ is a random error assumed to be independent for different measurements and having a log normal distribution with unit median. In terms of logarithms, $\epsilon^* = \ln \epsilon$ represents the additive error $\hat{U}^* - U^*$ and is a zero-mean normal random variable.

IV. Classification Criteria

This section assumes that just one measurement is made of each measured factor (modifications for repeated measurements are dealt with in section V). By combining these measurements in the same product–quotient as appears in the original definition of X, we obtain a *total measured* factor, $\hat{X}$. Example 1 (section II) would thus have $\hat{X} = \hat{U}\hat{W}$; and example 2, $\hat{X} = \hat{U}\hat{V}$.

For the remaining factors which occur in Y and are not measured, the product–quotient of their long-term geometric means (obtained from records) is simply taken to define the quantity ν_Y (which is also the long-term geometric mean of Y). Then the *calculated severity*, $\hat{S}$, is defined simply as

$$\hat{S} = \hat{X}\nu_Y/G \tag{4}$$

The classification procedure is based simply on $\hat{S}$ and is to classify the source as clean if $\hat{S}$ does not exceed a critical level, c, given below. Since random variation is involved in determining $\hat{S}$, there will always be some probability of misclassifying a source. Before c can be specified, the maximum acceptable probability β of misclassifying a dirty source must be specified. Then if, as above, α is the acceptable expected proportion of days on which the severity, S, exceeds 1, the classification procedure is as follows: If $\hat{S} > c$, classify the source as dirty; if $\hat{S} < c$, classify the source as clean. c is given by

$$c = \exp[-\sigma_0 z_{1-\alpha} + \sigma z_\beta] \tag{5}$$

where σ_0^2 = sum of temporal variances of *all* factors U^*, V^*, $W^*, \ldots$, σ^2 = sum of total (temporal and measurement-error) variances of all *measured* factors, $\hat{U}^*$, $\hat{V}^*$, $\hat{W}^*, \ldots$, and $z_{1-\alpha}$ and z_β are respectively the $1-\alpha$ and β percentage points of the standard normal distribution. For example, if $\alpha = 0.05$ and $\beta = 0.01$, then $z_{1-\alpha} = 1.65$ and $z_\beta = -2.33$.

To repeat, this classification procedure restricts the probability of misclassifying a dirty source, to the specified amount, β. The critical level, c, is chosen to satisfy that requirement. That is, in the long run, no more than a proportion, β, of dirty sources will be misclassified. (It is also of interest to determine probabilities of misclassifying *clean* sources; as already noted, this is taken up in section VI.) Finally, the basic assumptions include knowledge of the various temporal and error variances σ_0^2 and σ_e^2. It is envisaged that these will be reasonably well-known from previous experience with similar situations or can be relatively reliably estimated (an appropriate procedure being given in Appendix B). Where this is not the case, and estimates based on just a few current measurements are used, the theory should be used only with considerable caution.

V. Multiple Measurements

In practice, some factors would be measured more than once, giving, of course, an improvement in the procedure. Two points should be noted. First, different factors may be measured different numbers of times, and on one or more days. Second, repeated measurements on the same day will show less variation than measurements on different days, since the latter will include daily variability as well as measurement errors. The procedure obviously should reflect these facts. While more complicated models are possible, it is simply assumed here that repeated observations on a given day differ only by virtue of independent measurement or sampling errors (cf. Appendix A for the precise model).

In fact, the procedure calculating $\hat{S}$ requires very little modification. If one again assumes that $S = XY$ and $\hat{S} = \hat{X}\nu_Y$, where ν_Y is unaltered from the previous definition and $\hat{X}$ is the same product–quotient of the measured factors $\hat{U}, \hat{V}, \ldots$, the sole difference lies in the method of calculation of these factors, which are now not single observations but appropriate (weighted geometric) averages of observed values.

Specifically, suppose that U is measured on k days, with n_i measurements being made on the ith day. $\hat{U}_i^*$ will denote the (arithmetic) average of the n_i ln values on the ith day. Then the computed factor value, $\hat{U}$, to be used has the logarithm

$$\hat{U}^* = \sigma^2 \sum \alpha_i \hat{U}_i^* \tag{6}$$

i.e., a weighted average of the daily averages of the (logarithmic) values measured for U. The weights α_i are simply

$$\alpha_i = [\sigma_{U^*}{}^2 + (\sigma_e{}^2/n_i)]^{-1} \tag{7}$$

where $\sigma_{U^*}{}^2$ and $\sigma_e{}^2$ are respectively the temporal and measurement-error variances of U^* and $\hat{U}^* - U^*$, and $\sigma^2 = (\Sigma\alpha_i)^{-1}$. Since $\hat{U}^* = \ln \hat{U}$, the value of $\hat{U}$ may thereby be obtained to be combined with those of the other measured factors to give X, and hence $\hat{S} = \hat{X}\nu_Y$. The classification procedure is now exactly as before, i.e., to classify the source as clean or dirty according to $\hat{S} < c$ or $\hat{S} > c$ where again

$$c = \exp[-\sigma_0 z_{1-\alpha} + \sigma z_\beta]$$

$\sigma_0{}^2$ again being the sum of temporal variances of all factors and σ^2 being the sum of the values of $(\Sigma\alpha_i)^{-1}$ for each measured factor.

VI. Misclassification of Clean Sources

The above procedure was designed so that the probability of misclassifying a dirty source does not exceed a specified value, β. It is inevitable that a clean but almost dirty source may be misclassified with probability close to $1-\beta$. Since β will be small (e.g., 0.05), this means a high probability (e.g., 0.95) of misclassifying such a "nearly dirty" source.

Of course, misclassifications of such sources may not be too important since they are close to actually being dirty. However, it is important that sources which are rather clean not be classified as dirty with any great frequency. Hence, it is important to determine the probability of misclassifying such a clean source.

As a measure of the degree of cleanness of a source, the expected proportion, γ, of days in which the severity exceeds 1 is used. γ is referred to as a *source quality parameter*. For clean sources, of course, $\gamma \leq \alpha$, whereas $\gamma > \alpha$ for dirty sources. It is readily shown (as in Appendix A) that the probability of misclassifying a (clean) source with quality parameter, γ, is

$$1 - \Phi[z_\beta + (\sigma_0/\sigma)(z_{1-\gamma} - z_{1-\alpha})] \tag{8}$$

in which Φ denotes the standard normal distribution function, $z_{1-\gamma}$ is the $1-\gamma$ normal percentile, and σ_0 and σ have the same meanings as in section IV or, more generally, as in section V. Since, for given value of α and β, this is a function of γ depending on just one parameter σ_0/σ, a family of curves can be obtained for varying values of this parameter, such as those plotted in Figure 1, for $\alpha = 0.1$, $\beta = 0.05$, $\sigma_0/\sigma = 1, 1.5, 2$.

Note that the sources considered with $\gamma < 0.1$ are to be regarded as clean since $\alpha = 0.1$ has been chosen for this illustration.

Also note that, when $\gamma = \alpha$ (i.e., the source is on the clean–dirty borderline), the formula gives the value $1 - \Phi(z_\beta) = 1 - \beta$ for classifying the source as dirty and hence the value β for classifying the source as clean. This is consistent with the

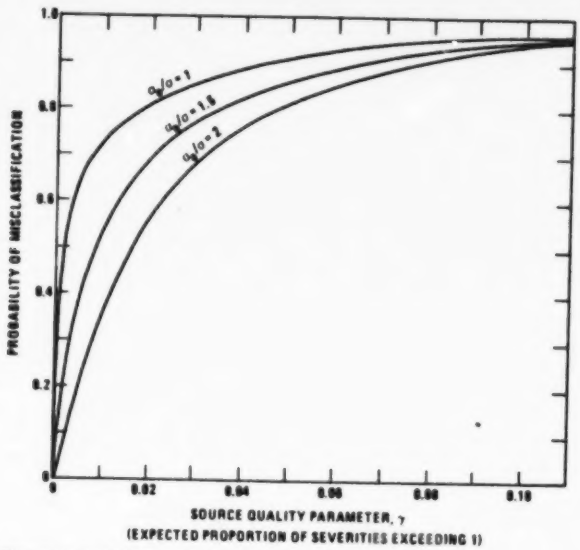

Figure 1. Probability of misclassifying clean sources.

fact that the probability of misclassifying a dirty source is at most β and, in fact, is exactly β for a dirty, borderline source. Further, if there is no temporal variability, $\sigma_0 = 0$ but $z_{1-\gamma} = \infty$ for a clean source. In that case the above formula for misclassification of a clean source with (constant) severity, S, is properly modified to read

$$1 - \Phi(z_\beta - S^*/\sigma)$$

Finally, note that these results are simply (1 minus) so-called power functions for the statistical test underlying our classification criterion.

VII. Applications

For illustration, data are used from a study (*4*) concerning the concentration of total suspended solids (TSS) in the effluent from a textile plant. Daily measurements were made over a 40-day period, with repetitions on some days to give a total of 57 observations. The study was therefore more extensive than might be generally undertaken, and does provide adequate variance estimates for our purposes. Specifically, the procedure of Appendix B provided the following values for the variance of $U^* = \ln$ (concentrations of TSS) and its measurement error: $\sigma_{U^*}^2$ = daily variance of $U^* = 0.081$; σ_e^2 = measurement-error variance of $\hat{U}^* - U^* = 1.33$. (In this study U is reported in mg/L though the variances of the logarithms are independent of units, since a change of units leads to an additive constant.) The severity at the source itself is considered first, and then an ambient case is discussed, based on discharge into a hypothetical stream.

Severity at the Source. In this case X is composed of just one factor U, the TSS concentration, and there is no nonmeasured factor Y. The critical level, c, depends on the number of observations used in computing S. The critical levels are given in Table I for 1 measurement, 10 measurements on 1 day, and 2 measurements on each of 5 days. The calculation uses a 5% allowable proportion of exceedances of unit severity for a clean source and a 5% maximum misclassification probability for dirty sources.

When the value $G = 368$ mg/L obtained from the Code of Federal Regulations (*5*) is used, an inspection of the data shows that on just 10 of the 57 cases was the severity less than 0.088. Hence, based on a single measurement, the conclusion in 47 of the 57 cases would have been that the source is dirty. However, selecting 5 (representative) days on which two

Table I. Critical Severity Levels

	1 measurement	10 measurements on 1 day	2 measurements on each of 5 days
σ_0^2	0.081	0.081	0.081
σ^2	1.41	0.214	0.149
c	0.088	0.29	0.33

measurements were made, we have the following measured concentrations (in mg/L): 60, 28; 36, 46; 72, 40; 88, 100; and 1540, 64.

The calculated severity is $^{76}/_{368} = 0.21$, which is less than the critical level of 0.33. Hence, the repeated sampling gives the more reliable conclusion that the source is in fact clean. This difference in conclusion illustrates the decrease in misclassification probability of a clean source with increasing numbers of measurements.

It is also interesting that the measured daily severity exceeds 1 in 9 of the 57 cases. This would suggest a rate of 15% or so, rather than the maximum 5% required of a clean source. However, many of these exceedances will be caused simply by measurement errors (the variance of such errors being large compared with the temporal variance), and the actual severity will not necessarily exceed 1 in those cases. This highlights the importance of distinguishing between the two sources of variation and not drawing conclusions based purely on a superficial examination of the data.

Ambient Severity. Suppose that the source above discharges into a stream flowing at 50 000 gal/min (3 m^3/s). The formula of example 2 (Section II) is used:

$$C = UV/W$$

applicable after complete mixing, where U = concentration of pollutant in effluent, measured as above (in mg/L), U^* having temporal and measurement-error variances of 0.081 and 1.33 respectively, V = effluent discharge rate, measured (in m^3/s), V^* having temporal and measurement-error variances of 0.1 and 0.05, respectively, and W = river flow rate (m^3/s) is not measured but it is known that W^* has mean 1.1 and temporal variance of 0.1.

Based on these assumptions:

$$\sigma_0^2 = 0.1 + 0.1 + 0.081 = 0.281$$

$$\sigma^2 = 0.1 + 0.081 + 1.33 + 0.05 = 1.561$$

so that (for one measurement of each of U and V, and $\alpha = \beta = 0.05$) $c = 0.053$.

For two measurements of U and V on each 5 days, $\sigma_0^2 = 0.281$, as above

$$\sigma^2 = \tfrac{1}{5}(0.081 + 1.33/2) + \tfrac{1}{5}(0.1 + 0.05/2) = 0.174$$

giving $c = 0.21$.

The calculated severity is

$$\hat{S} = \hat{U}\hat{V}/(\nu_w G) \qquad (9)$$

where U and V are calculated from the observations and ν_w is the long-term geometric mean of W, obtained from past data. Assuming that U and V are each measured twice on 5 days with the V values being as in Example 1 above, $\hat{U} = 76$ mg/L. If $\hat{V}$ is correspondingly measured and found to be 0.0165 m^3/s (280 gal/min—approximately so for the plant considered in ref 4, and ν_w and 3 m^3/s, then

$$\hat{S} = 0.42/G$$

In this case the water standard from ref 5 is used for G (viz., 25 mg of TSS/L) to obtain

$$\hat{S} = 0.017$$

which is well below the critical level $c = 0.21$ for ambient severity. Of course, this figure is highly dependent on the assumed stream flow rate and would be much larger for streams where the flow rate is smaller.

In concluding, it should be noted that the classification problem has been recognized as a statistical hypothesis testing situation in the past, and various approaches have been used to obtain severity distributions from measurements on factors (cf. ref 6). The log normal assumptions used here should be appropriate (borne out by tests of the data used), and the multiplicative models are natural and recognized as at least reasonable approximations. It has been assumed throughout that appropriate variances are known from experience or can be reliably estimated. Obviously, the theory would not apply so well if those variances must be estimated from very small amounts of data. It may also be desirable to take account of day-to-day correlations which are sometimes found to be present in such data, for example, the averaging effects of settling ponds. Finally, no claim is made here (any more than for any applied statistical analysis) that the underlying assumptions hold exactly. But it is felt that a statistical analysis, based on reasonable assumptions, is much preferable to the use of highly conservative safety factors and worst-case considerations which are likely while recognizing bad sources, to misclassify the vast majority of clean sources, thereby causing considerable unnecessary expense.

Acknowledgment

We are greatly indebted to our colleagues W. D. Baasel and K. E. Rowe, of the Special Studies Staff, Industrial Environmental Research Laboratory. It is a pleasure to express our sincere appreciation to them for extensive conversations, as well as data location and computational advice and help.

Appendix A

Mathematical Details. This appendix indicates the details of the development of the classification procedure. As noted previously, the severity S is assumed to have the multiplicative form $S = XY/G$ or, in terms of (natural) logarithms

$$S^* = X^* + Y^* - G^* \tag{10}$$

where X is a product–quotient of independent log normal factors to be measured, and Y is log normal with known parameters and hence will not be measured.

Since the expected proportion of days on which $S > 1$ is $P(S > 1)$, the classification problem is equivalent to the problem of testing the hypothesis

$$H_0: \quad P(S > 1) > \alpha$$

Now $P(S > 1) = P(S^* > 0) = 1 - \Phi(-\mu/\sigma_0)$, where Φ denotes the standard normal distribution functions, $\mu = \mathrm{E}(S^*)$, the mean of S^*, $\sigma_0^2 = \mathrm{var}\, S^*$, the temporal variance of S^* (i.e., the sum of temporal variances of all of the factors $U^*, V^*, \ldots$). Hence H_0 may be written as

$$H_0: \quad \mu > -\sigma_0 z_{1-\alpha}$$

where, as previously defined, $z_{1-\alpha}$ is the $1 - \alpha$ percentile of Φ.

Suppose, for simplicity of notation, that X consists of just the factor U, which is measured n_i times on the ith of k days. If $\hat{U}_{ij}$ denotes the jth of the measurements on day i, we assume the model

$$\hat{U}_{ij}^* = U_i^* + \epsilon_{ij} \qquad 1 \le j \le n_i, 1 \le i \le k \tag{11}$$

where U_i^* is the true U^* value on the ith day and ϵ_{ij} values are independent normal observation errors with zero means and variances σ_e^2.

Let $\hat{U}_{i\cdot}^*$ denote the average of the n_i values of $\hat{U}_{ij}^*$ as measured on the ith day and define

$$\hat{U}^* = \sigma^2 \sum_{i=1}^{k} \alpha_i \hat{U}_{i\cdot}^* \tag{12}$$

where the α_i values are weights designed to minimize the variance of this linear combination, viz.

$$\alpha_i = (\sigma_{U^*}^2 + \sigma_e^2/n_i)^{-1} \tag{13}$$

(in which $\sigma_{U^*}^2$ is the variance of U^*) and $\sigma^2 = (\Sigma_{i=1}^{k} \alpha_i)^{-1}$. It is readily checked that $\hat{U}^*$ has the same mean, μ_{U^*}, as U^* and (from the definition of the α_i) variance σ^2.

By definition, the calculated severity $\hat{S} = X\nu_Y/G$ so that

$$\hat{S}^* = \hat{X}^* + \mu_{Y^*} - G^* = \hat{U}^* + \mu_{Y^*} - G^* \tag{14}$$

where μ_{Y^*} is the mean of Y^*. But since $S^* = U^* + Y^* - G^*$, it follows at once that $\hat{S}^*$ has mean $\mu_{X^*} + \mu_{Y^*} - G^* = \mu$, the mean of S^*, and variance σ^2. Thus, the hypothesis H_0 to be tested may be rewritten as

$$H_0: \quad \mathrm{E}(\hat{S}^*) > -\sigma_0 z_{1-\alpha}$$

The appropriate test of H_0 is to reject it (classify the source as *clean*) if $\hat{S}^* < c_0$ where

$$\beta = P(\hat{S}^* < c_0 | \mu = -\sigma_0 z_{1-\alpha}) = \Phi((c_0 + \sigma_0 z_{1-\alpha})/\sigma) \tag{15}$$

giving

$$c_0 = -\sigma_0 z_{1-\alpha} + \sigma z_\beta$$

Translated back into original units the test thus takes the form given in sections IV and V; i.e., classify the source as clean if

$$\hat{S} < c = \exp[-\sigma_0 z_{1-\alpha} + \sigma z_\beta]$$

The above derivation concerns the case where X consists of a single factor U. When X is a more general product–quotient of factors, σ_0^2 is still the sum of the temporal variances of all factors $U^*, V^*, \ldots$ and σ^2 is obtained as the sum of the values of $(\Sigma\alpha_i)^{-1}$ for all of the *measured* factors.

Finally, the probability of correctly classifying a clean source is simply obtained as a power function for this test. If the source is such that $P(S > 1) = \gamma < \alpha$, it follows that $\gamma = P(S^* > 0) = 1 - \Phi(-\mu/\sigma_0)$ so that $\mu = -\sigma_0 z_{1-\gamma}$. The probability of correct classification of this clean source is thus

$$P(\hat{S}^* < -\sigma_0 z_{1-\alpha} + \sigma z_\beta | \mu = -\sigma_0 z_{1-\gamma}) = \Phi(z_\beta + (\sigma_0/\sigma)(z_{1-\gamma} - z_{1-\alpha})) \tag{16}$$

as stated in section VI.

Appendix B

Estimation of Variances. As noted, if the required variances σ_0^2, and σ_e^2 are unknown for one or more of the measured factors, their estimates from the data may be used, provided these are based on sufficiently many observations to yield close approximations. If, as above, the factor U is measured n_i times on the ith of k days, $\hat{U}_{ij}$ being the jth of these measurements, then the appropriate estimates $\hat{\sigma}_e^2$ and $\hat{\sigma}_0^2$ of σ_e^2 and σ_0^2 are

$$\hat{\sigma}_e^2 = \frac{1}{M - k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\hat{S}_{ij}^* - \hat{S}_{i\cdot}^*)^2 \tag{17}$$

$$\hat{\sigma}_0^2 = \frac{1}{M - M^{-1}\sum n_j^2} \times \left\{ \sum_j n_j (\hat{S}_{j\cdot}^* - \hat{S}_{\cdot\cdot}^*)^2 - (k - 1)\hat{\sigma}_e^2 \right\} \tag{18}$$

where $M = \Sigma n_j$ is the total number of observations and the

dots (in place of subscripts) denote averages over those subscripts.

Nomenclature

C = concentration of a discharged substance
G = standard or goal level for C
S = severity, C/G
α = allowable expected proportion of exceedances of unit severity
$\hat{S}$ = calculated severity of source
c = critical level of $\hat{S}$ for classification of source as "clean" or "dirty"
β = maximum probability for misclassifying a dirty source
U, V, W, Z = source and ambient factors in model for C
X, Y = combined (product–quotients of) measured and nonmeasured factors, respectively, in model for C
ν_Y = long-term geometric mean of Y
$\sigma_{U^*}^2, \sigma_e^2$ = temporal and measurement-error variances of a factor $U^* = \ln U$
σ_0^2 = sum of temporal variances of all (log-transformed) factors
σ^2 = sum of total (temporal and measurement-error) variances of all measured (log-transformed) factors (appropriately weighted in the case of repeated measurements)
n_i, k = number of measurements of a factor on ith day, $1 \le i \le k$
α_i = weights for daily average of a measured factor
z_γ = γ percentage point of the standard normal distribution
* = natural logarithm ($S^* = \ln S$, etc.)
. = when used in place of a subscript indicates the average over that subscript, as in $\hat{U}_{i\cdot}^* = \text{av}_j\, \hat{U}_{ij}^*$

Literature Cited

(1) Baasel, W. D.; McAllister, R. A.; Kingsbury, G. L. *Chem. Eng. Prog.* **1980**, *76*, 37–51.
(2) Hangebrauck, R. P.; Denny, D. A.; Tucker, W. G. Research Triangle Park, NC, Aug 1979, "Nomenclature for Environmental Assessment Projects. Part 1. Terminology for Environmental Impact Analysis," U.S. EPA, Industrial Environmental Research Laboratory memorandum.
(3) "Guideline for the Interpretation of Ozone Air Quality Standards," U.S. Environmental Protection Agency, Jan 1979, EPA-450/4-79-003 (NTIS PB 292271).
(4) "Textile Mill Report, Plant D," Engineering Science, Inc., Atlanta, GA, Aug 1978, Alpha 3 Report on EPA Grant R-804329.
(5) "Quality Criteria for Water," U.S. Environmental Protection Agency, July 1976, EPA-440/9-76-023 (NTIS PB 263043).
(6) Serth, R. W.; Hughes, T. W.; Opferkuch, R. E.; Eimutis, E. C. Dayton, OH, Aug 1978, "Source Assessment: Analysis of Uncertainty—Principles and Applications," EPA Report EPA-600/2-78-004u (NTIS PB 80-131 485).

Received for review October 27, 1980. Accepted June 16, 1981.